DOCUMENT RESUME

ED 458 260 TM 033 423

AUTHOR Lewis, Anne
TITLE Benchmarks for Accountability: Are We There Yet? 1999 CRESST Conference Proceedings (Los Angeles, California, September 16-17, 1999). CSE Technical Report 547.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-547
PUB DATE 2001-09-00
NOTE 51p.
CONTRACT R305B960002
AVAILABLE FROM For full text: http://www.cse.ucla.edu.
PUB TYPE Collected Works - Proceedings (021)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Accountability; *Benchmarking; Conferences; Educational Quality; Educational Technology; Elementary Secondary Education; *High Stakes Tests; *Standards; *State Programs; Urban Schools
IDENTIFIERS *Center for Research on Eval Standards Stu Test CA

ABSTRACT

This report outlines the major themes of the 1999 annual conference of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), which focused on creating benchmarks for accountability in education. Opening the conference, a panel of experts focused on recent work regarding standards, quality, and fairness in the assessment field, with a summary of CRESST's work and a discussion of accountability issues presented by Eva L. Baker. The national context for assessment was discussed by Joseph Conaty, director of the National Institute on Student Achievement, Curriculum, and Assessment. As part of this panel, Richard Elmore reviewed the recent report on Title I Testing. Another session focused on testing and accountability for high stakes purposes, featuring presentations by Lorraine McDonnell, Edward Haertel, and Lorrie Shepard among others. The third panel discussion discussed the validity of testing and assessment of all children, and another panel considered the impact of accountability systems on instruction and outcomes. Small group discussions, in concurrent sessions held each day of the conference, focused on: (1) technology use; (2) the implementation of assessment in large urban school systems; (3) reporting school and district progress; (4) using multiple measures of assessment; (5) measuring school performance through the California Academic Performance Index; (6) issues surrounding college entrance; and (7) systemic and classroom perspectives on the alignment of goals, instruction, and the use of assessment. In the closing session, Robert Mislevy discussed making sense of the data from assessment systems, and other presenters continued with the discussion of testing issues and validity concerns. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 458 260

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

K Hurst

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

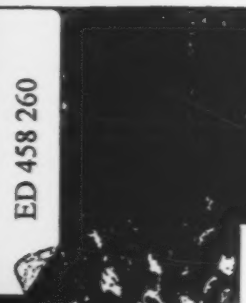
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



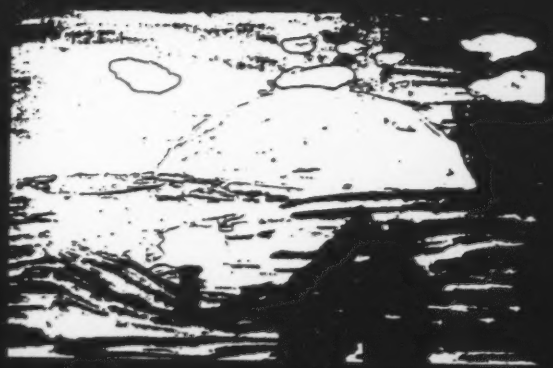
1999 CRESST Conference Proceedings

Benchmarks for Accountability: Are We There Yet?

CSE Technical Report 547

Anne Lewis

TM033423



UCLA Center for the Study of Evaluation

BEST COPY AVAILABLE

1999 CRESST Conference Proceedings

Benchmarks for Accountability: Are We There Yet?

CSE Technical Report 547

Anne Lewis

September 2001

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2001 The Regents of the University of California

The work reported herein was supported in part by the Educational Research and Development Centers Program, PR Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or polices of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

1999 CRESST CONFERENCE PROCEEDINGS

BENCHMARKS FOR ACCOUNTABILITY: ARE WE THERE YET?

Anne Lewis

The work to create good benchmarks for accountability in education proceeds at almost a furious pace, as papers and discussions at the 1999 CRESST National Conference affirmed over and over again. Most heartening is the greater understanding and use of accountability at the classroom level. Yet, efforts to maintain continuity and to establish well-researched standards for accountability in a volatile political and policymaking environment are struggling—both to keep up and to be heard.

It was in this context—of endeavoring to make assessment work under challenging conditions—that the 300 participants at the conference discussed often controversial policy aspects of accountability, primarily at national and state levels. Moreover, they learned from researchers about their ongoing studies and assistance at classroom and school levels to make accountability tools useful and integrated into instruction.

Opening the conference, a panel of experts focused on recent work regarding standards, quality, and fairness in the assessment field. Interest in standards and accountability is expanding, said Eva Baker, co-director of CRESST, in a presentation also representing CRESST Co-director Robert Linn, because they affect education at four levels. Children, teachers, and schools—"the people we do this stuff to"—represent level one. Next, standards for accountability developed by the Joint Committee (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999) should be the essence of testing instruments. This guide on the limits of and uses for tests has just been published. At the third level of interest in standards and accountability are the practices that have widespread effect, such as the standards developed by the Office for Civil Rights and evaluation standards for Title I. At the final level is the need to assess accountability systems themselves.

Education accountability, Baker noted, is a drama in one act. As the plot goes, new leadership arrives on the horse of "higher standards." The new leadership

unveils an accountability approach to remedy existing defects, this system is debated and then enacted, participants comply but legal and technical entanglements arrive, and the horse is replaced by new leadership that looks much the same. "As an audience," she said, "we suspend disbelief each time the system changes despite the fact that change is historically inevitable. Because politicians want to make their mark and new technologies come along, the implementation rarely is allowed to develop.

To avoid perpetual surprise and dismay, "we must not be shocked every time" by these changes, Baker said. Instead, change and flux should be anticipated, but there also need to be clear criteria for the review and accomplishments of accountability systems. Moreover, the media and other influences should set public expectation that policymakers have both the responsibility and authority to use the criteria. Finally, those who want accountability to succeed must work hard to provide continuity "when what we have is a discontinuous system."

This means "watching the watchers" to make sure they adhere to a set of accountability standards that have been agreed to and disseminated widely, and to which people feel they need to pay attention. The standards for accountability systems must be built on principles that include validity, fairness, credibility, and utility. Drawing analogies from the revised *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), Baker said tests must

- state purpose(s) and minimize negative consequences;
- give evidence of technical quality for each purpose;
- document relationship to content standards;
- match instruction with test content if the test is to be used for high stakes (promotion);
- give evidence of suitability of the test for the program and for the test population;
- provide a basis and evidence for expectations when use of a test or system implies a specific outcome;
- minimize possible misinterpretations of data with appropriate context;

- make no student decision on the basis of one test;
- make sure test preparation does not adversely impact the validity of results;
- include in reports classification error and error in measurement of change; and
- use trained personnel for public interpretation.

Accountability systems need to emphasize their objective achievement of such goals as improving student access and performance, improving the quality of schools, managing resources, raising public confidence, and building the long-term health of the system. They should abide by criteria that will assure quality systems. Baker and Linn propose that such criteria include

- articulation of goals and purposes to be achieved;
- evaluation over time of the accountability system's effects on its goals;
- design and documentation of valid, interpretable indicators for particular goals or purposes;
- planning of symmetrical accountability for institutions, individuals, policy, management, teachers, and students;
- matching of philosophy of the system to the indicators;
- making the combination of multiple measures intuitively clear;
- making sure the indicators are not in conflict, or that conflicts can be resolved;
- assuring that progress monitoring assesses change in quality; and
- basing rewards on efforts of schools (institutions) with different challenges.

It is not enough to add all the ingredients of a good accountability system, stir, and serve, Baker said. Not all components of a system will work well or coherently support the validity of all goals, nor will all components show early impact. Moreover, the "watchers" need a development perspective, focusing on patterns of effects. And they should shift the focus of attention to those indicators that show "what schools can do and are doing."

To assure high standards for accountability systems, Baker outlined a research agenda that includes validity of measures and indicators; psychometric and statistical properties; reporting paradigms, inferences, and impact of accountability

results on children, schools, and systems; developmental patterns of indicator impact; new, scalable measures of classroom practice; evidence of alignment at an operational and conceptual level; creating indicators that tap major goals (e.g., reduction of the impact of background characteristics on performance); instructionally based change; and validity of measures for all children.

National Context for Assessment

Making standards for accountability concrete, Joseph Conaty, director of the National Institute on Student Achievement, Curriculum, and Assessment in the U.S. Department of Education,[1] reminded participants why tests and assessment are key issues in implementing education reforms. The latest National Assessment of Educational Progress results, he said, "show once again that 68% of children in high-poverty schools fail to read at the basic level. All of the conversation on accountability should be played out against a background of its effect on children's performance."

Presenting an historical review, Conaty said that in the early 1990s, states began to focus on creating a standards-based system based on high-level content linked to performance standards and assessments, and national leadership followed. In 1994, Goals 2000 helped states set challenging content and rigorous performance standards, while a key requirement in the reauthorization of the Elementary and Secondary Education Act concerned standards and assessment for Title I.

The conversation at the national level, according to Conaty, "has reached a second generation now that most states have come to some sense of agreement on what content standards mean for their states." They are seeking answers to the question of how to measure and give children opportunities to reach standards. This second phase "has raised a set of issues around accountability and federalism that is playing out in every district and state and at the national level."

Conaty warned against the development of multiple systems of accountability that are contradictory. "You have to keep in mind what standards-based reform is all about," he said. "There is a large body of evidence in reading, for example, that indicates that with a variety of high-quality programs and interventions, we can reduce those who can't read to about 1% in high-poverty districts." The key is to

[1] Joseph Conaty is currently the director of the Special Initiatives Unit in the Office of Elementary and Secondary Education, U.S. Department of Education.

align the curriculum, assessment, and professional development "and hold people responsible for results," which, he said, are issues that will be the subject of a number of upcoming reports. Furthermore, to avoid conflicts, the state and national levels must have a shared agreement on what constitutes a high-quality assessment system.

At the national level, Title I assessment is a major focus because states were required to have content and performance standards by 1998 and final assessment systems aligned to the standards by 2000. School-level and district-level accountability systems are to be aligned to standards, using assessments and tests as one component. He pointed out that "it is important to distinguish between accountability systems and assessments as a tool within those systems." Furthermore, there is debate over civil rights and individual testing. If high-stakes testing produces disparate impact by race, sex, or origin, "the impact may be illegal, or permissible if the tests are valid for the purposes for which they are used and if no practical assessment alternative produces less of a disparate impact."

From a national perspective, the unit of analysis in Title I is schools and districts. The goal is to integrate Title I accountability with state accountability systems to use the same data and information to drive improvement in classroom instruction.

Conaty said these three requirements—Title I, civil rights, and accountability—represent the "big picture." A quality accountability system recognizes the difference between institutional accountability and accountability for decisions made about individual students. At the state and district levels, tests are a necessary component of accountability systems, but there are other indicators that can be considered, such as graduation and dropout rates. He emphasized that Title I does not require high-stakes individual testing.

Certain technical issues, such as what constitutes technical quality, cut across all the conversations about assessment, Conaty said. Also, the Department [of Education] and states have been in discussions on three components of an accountability system:

- **Alignment.** Assessments must measure what students are expected to learn. The challenge is to design tests that adequately measure content of instructional practice and that are not so overwhelmed by background effects that they are not capturing true school effects. This arises as a

problem when using or retrofitting off-the-shelf tests or when using a single test to cover several content domain performance levels.

- **Multiple measures.** Through use of multiple measures, issues develop over inferences of student performance and content domains, the validity of alternative forms, and their combination.

- **Inclusion.** The purpose of accountability systems is to be sure the same expectations hold for all children, and that all children reach high standards. The technical quality of accommodations and modifications must be assured, as well as alternative assessments for children with different needs and the integration of accommodations into the accountability system.

"If you keep these three issues in mind," Conaty concluded, "you can see a way through the major reports that will be coming out on assessment and see assessment as a component of larger accountability systems."

Title I Testing Report

Just a few days after *Testing, Teaching, and Learning: A Guide for States and School Districts* (1999) was released by the National Research Council, one of its co-editors, Richard Elmore, told the CRESST conference that the core of the report focused on testing as a tool for education improvement. Produced by the Committee on Title I Testing and Assessment, which Elmore chaired, the report includes much that is specific to Title I, but its main elements raise more general issues about the purposes of standards-based reforms and the role of assessment in them.

Describing the policy context, Elmore noted that "there is a tendency for people to assign some sort of naive directionality to standards-based and performance-based accountability reforms, that someone out there has designed an accountability system predicated on the assumption [that] teachers are inadequate and schools are failing and the best solution is a hard-nosed one." Not true, he said. "What is really happening is that there is no directionality and multiple directionality. This movement . . . is primarily rooted in state politics, not federal or local politics, and it is extremely resilient and durable."

In Elmore's opinion, the so-called accountability movement "isn't going anywhere except straight ahead. You can get a good or a bad version, but you can't choose whether to have it or not." The push for accountability springs from state economies, he said. Rather than a rational pattern that follows arrows from the local, to the state, to the federal level, the accountability movement is more like a puzzle

with interrelated pieces. If there is any directionality at all, it is in using accountability more than ever before to "open up what gets taught and with what consequences."

With regard to Title I, all effects of federal policy are interrelated, between policies and requirements and the contexts and capacities of states/localities. As the report emphasizes, it is impossible to address accountability without also talking about instructional improvement. "It is deceptive and misleading," Elmore said, "to suggest that you can solve problems in schools by designing and implementing good accountability systems. If they are not deeply rooted to instruction, they are not likely to have a constructive effect."

The report also focuses heavily on the problems presented by limited English proficient students and students with disabilities. "The job of an accountability system is to create monitoring, data, and clear criteria to bring as many students as possible into assessment systems," with an overriding concern that the measures be valid assessments of what students are learning. Also, the report calls for the inclusion of K-3 students in assessment systems, "but not with norm-referenced tests." Although the Title I testing committee was not asked to study professional development, the logic of the report, Elmore said, extends to creating strong professional development within an accountability system.

A fundamental theory behind the report's thrust, Elmore concluded, is that "schools cannot respond to external accountability systems unless there are internal accountability systems." School organizations must be healthy, monitoring their own performance as a matter of routine, with the district role being one of helping schools to monitor the quality of instruction and to develop their own assessments.

Response to the Panel

Reviewing the major points made by presenters, Sylvia Johnson of Howard University noted that technical improvements in assessments made in recent years "may have been oversold as engines of school improvement." They have an important but not primary role, in her opinion, and are most effective when they can provide information on what needs to be done to improve education for all children. An expectation of positive performance is the "unspoken context" of the Title I testing report, which means "we have a conditional model, with a positive expectancy as the major condition in terms of school improvement." Assessment is useful primarily to the extent that it contributes to student learning and includes an

emphasis on quality instruction and optimal expectations of academic success. Without such expectations, an education improvement system "becomes a series of assessment reports and recriminations without real progress for districts, schools, teachers, or students," Johnson said.

Asked about the possibility that differences in capacities and will among the states might lead to even greater variability in the performance of children under a standards-based system, Conaty replied there is evidence that implementing such reforms improves the performance of all children. For example, in Texas and North Carolina, he said, the gains are not being achieved solely by moving children at the top higher but by moving children at the lower ends of the performance distribution to higher levels. The federal level is seeking partnerships with states, he added, but there will be times when "we will say there are financial consequences for districts that continue to be low performing."

Testing and Accountability for High-Stakes Purposes

Using tests for high-stakes purposes takes assessment beyond technical issues and into the thorny realm of politics. Whereas Eva Baker's presentation focused on what good thinking needs to go into assessment and how smart people should respond to accountability policies, Lorraine McDonnell of CRESST/University of California, Santa Barbara, presented high-stakes testing from the policymakers' viewpoint.

Assessment policy now is intended to be more than just an accountability tool, she said. The policy theory behind high-stakes testing makes several assumptions: Testing can reach into classrooms and change instruction; information alone is insufficient to motivate educators to teach well; and education needs rewards and sanctions.

As for the political context, a lot of survey data show that the public favors high-stakes testing "while also seeming to acknowledge its limitations," McDonnell said. There are no significant differences by race or parental status as to these opinions, and teachers also favor high-stakes testing. The public is willing to accept some of the negative consequences associated with high-stakes testing such as higher failure rates. The findings are consistent over time, and public opinion is continuing to play the role of legitimizing the tests, she said.

There also is a broad-based consensus among political elites regarding high-stakes testing. Both Democrats and Republicans at the national and state levels generally support it as a remedy to shortcomings of other kinds of accountability, though for different reasons. Some see it as an alternative to vouchers; some see it as a way to extend the scope of accountability. Some contend that if tests are not high stakes, students will not pay attention to them.

In this context of public approval, "politicians are making a decision to use the open window to push through policies before many would say these policies are ready to be implemented," McDonnell pointed out. "They see that the public wants it and the elite want it, so they are trading off political feasibility for implementation feasibility." Because the push for high-stakes testing is moving so fast, "we are sacrificing implementation feasibility."

Technical capacity will need to catch up with policy expectations, McDonnell advised. "Politicians say they understand the problems associated with high-stakes testing, but they see them as ones that can be fixed as we go along, as in Chicago," she said, emphasizing that technical advances are going to have to speed up.

The situation creates fewer opportunities for the policy critic, but more opportunities "for the fixer," who must be committed to the enterprise and emphasize concrete ways to ensure appropriate test use and adequate learning opportunities.

This political context is not fixed in stone, McDonnell observed, because "we don't know yet if the public is willing to tolerate true high stakes." Although public opinion endorses it, there are a growing number of examples where the bar has been lowered as people see the real consequences.

One of the political decisions under consideration is a technical minefield—determining policy on school comparability. Ed Haertel of Stanford University noted the issue grows out of a desire to level the playing field in order to make a fair and equitable assessment of schools, but all of the methods used in comparability share certain limitations. The methods for comparing schools should follow the purpose for doing it. The methods he discussed included selecting—finding a subset like the target schools—versus adjusting—producing a predictive achievement level for each school. With one method, a given school is matched with others in terms of students served or the kind of community, as contrasted to the other method, which calls for matching on predicted scores, or

comparing schools with similar predicted outcomes. The latter would have more statistical evidence than the former, Haertel said.

Another issue to consider is statistical versus substantive explanatory variables. Skin color, for example, is not a statistical variable. Haertel also discussed the desirable properties of matching schemes. Symmetry means that if school B is on school A's list of comparable schools, then A should be on B's. Transivity means that if school A is similar to school B, and B is similar to school C, then A is similar to C. If a fixed pool size is used, then certain schools will not totally fit, which may not be as much of a problem if the method is seeking a consistent degree of similarity instead of a fixed pool size. The matching schemes also should have intuitive appeal, transparency, and apparent fairness.

Haertel focused his discussion of contextual variables on California's list, but noted that such lists vary from one state to another. Taken into consideration in California are pupil mobility, ethnicity and socio-economic status; the percentages of teachers fully credentialed or emergency credentialed, and of pupils who are English language learners; the average class size; and whether the school is year round. Other possible contextual variables, according to Haertel, are school size, urbanization, grade range, percentage of pupils tested, per-pupil expenditure, average daily attendance, and hours of TV viewing.

Iowa probably is the only state that continues to use no adjustment for schools, putting everyone into the pool. Other illustrative methods discussed by Haertel include

- **comparison to predicted achievement.** This uses the logic of "effective schools research" and multiple linear regression. In this method it is important to include only "nonmodifiable" variables, "otherwise, there is a problem of adjusting away real school effects," Haertel said. For example, using TV as a factor would be misleading because "we know students watch TV and do homework at the same time." The potential problems with this method are that predictors could be incomplete or poorly measured, and comparisons could be made to patently different schools.

- **stratification methods.** This focuses on school characteristics rather than predicted outcomes. Choices must be made between a statistical or judgmental classification (for example, between a suburban classification and a socio-economic one) and between unidimensional and multidimensional schemes. The potential problems with this method include the heterogeneity of the educating challenge within clusters,

instability for schools near boundaries, and such tradeoffs as number versus size of strata.

- **floating comparison bands.** This solves the boundary problem by using fixed strata. The method uses prediction and rank, defining a group with the target school in the middle, using a regression equation. The problems include depending on heterogeneity as a function of predicted achievement ("a small change in achievement changes the bands"), the fact that the comparison is being made among patently different schools, and that schools at the top and bottom cannot be at the center of the band.

- **campus comparison groups.** This is used in Texas, where 100 schools are grouped based on dominant characteristics, and then narrowed successively to 40 schools based on race, percentages of ethnicity, economic disadvantage, LEP, and mobility. The potential problems with this method include different degrees of heterogeneity, suboptimal selection, and a requirement that assumes a common metric for all variables.

- **proximate schools in multidimensional space.** This was used in 1993 with the CLAS in California. It creates a comparable group defined by Mahalanobis distance. Characteristics were the mean socio-economic status and percentage of LEP, mobility, and percentage receiving welfare (AFDC). The potential problems include the difficulty in explaining the method and unstable weights with collinear predictors.

- **fixed region method.** This is a work in progress. The idea is to find a fixed-range definition of "similar" for each characteristic. If it works, it would be both transitional and symmetric, Haertel said. The problems are that it may not work, it is likely to have to relax criteria for schools at the extremes, and it depends on unequal comparison group sizes.

In summary, said Haertel, "there is no ideal method." There are tradeoffs or choices between selecting and adjusting, matching on characteristics and predicted outcomes, and stratification and customized comparison groups. However, the choices "are never just a technical or statistical matter," he said. "Whatever is done will be based on state legislative and political history." This means that fairness will remain elusive, and the methods will be limited by the quality of the test and the strength of empirical and theoretical support for actual uses and interpretations of test data. Using comparisons probably is better than doing nothing, Haertel concluded, "but it will lull people into a sense of doing something we really don't know how to take care of."

The high-stakes environment today also has stirred policy (and political) actions from coast to coast on retention, or "putting an end to social promotion." The issue illustrates earlier references to the "horse that comes charging in" bearing a

banner of reform, according to Lorrie Shepard of CRESST/University of Colorado at Boulder, but proponents never see, or care to see, the failure of similar retention policies in the past.

With some frustration, Shepard noted the seriousness of the actions because of the number of children who are going to be affected by them and the refusal to consider the research that documents the failure of retention as a policy intended to raise achievement. "Research on retention summarized by the National Academy of Sciences gets the same number of words in the media as the teacher who anecdotally 'has seen it work'," Shepard said.

Without equivocation, she said, comparative research studies show that, on average, repeating a grade either harms achievement or does not improve achievement in the years following the repeated grade itself. She based this statement on 63 studies. Students who have repeated a grade are about .38 standard deviation points behind those who were promoted. One study that found positive results from retention, by Carl Alexander, "was confused over vertical scale scores," Shepard said. Using different statistical analyses, the same data show that the performance of retainees went up in the repeat year, but subsequently dropped off.

A second major reason to question retention policies is that they increase the probability that students will drop out of school. Depending on the study, those who have been retained are 17% to 30% more likely to drop out than equally low achievers who were not retained. Once a student has been retained twice, there is a 90% likelihood the student will drop out. To offset the negative effect of being retained, Shepard explained, a student must have a 2.5 grade level increase in performance in the retained year, according to a study conducted a decade ago in Chicago.

Shepard also commented that political promises to end social promotion are made without knowledge of current retention policies. The number of elementary school students who are average increases each year so that by the time students reach the middle grades, 20% already have been retained. In urban districts such as Baltimore, more than 50% of students have repeated at least one grade by the time they enter middle school. These high retention rates existed before current efforts to end social promotion.

Nor have policymakers given thought to the problems of trying to end social promotion and raise standards at the same time. Projections of new requirements

might mean that large urban districts will be retaining 50% of their students in a single year. Based on the NAEP reading assessment, 40% of fourth graders are below the basic level, and the possibility of such failures "will test the will" of policymakers or lead to a lowering of standards, she said.

Retentions cost, on average, $5,500 per student, yet "all of the other things that work better cost less and have a better chance of succeeding," according to Shepard. Summer school, for example, has shown positive effects, although unfortunately in the current high-stakes climate, its emphasis has shifted to test preparation. The single most important thing that can be done to prevent retentions is to teach students to read well, early, "and stay with it, instead of waiting until they are in the fourth grade." Other fruitful alternatives are tutoring and before- and after-school programs.

Shepard also countered claims by some policymakers, educators, and lay citizens that the threat of flunking a grade will act as an incentive to make students work harder. No data support this, she said, and it is not known what proportion of low-achieving students do poorly because they are lazy versus those who do poorly because they have received poor instruction. If it is the latter case, "how is the threat of retention presumed to work?" she asked.

The discussion of high-stakes decisions had a unifying theme of symbolic politics, according to the discussant, Bella Rosenberg, of the American Federation of Teachers. Symbolic politics leads to pressure "to do real things," she said. The no-social-promotion policy adopted in Chicago, for example, was a political response to show the public that the new administration of the schools was serious about reforming the schools. Symbolic politics requires plans, policies, and technical responses to be in place so that the policy decisions can be turned into sensible practice and policy. As Lorraine McDonnell, pointed out, "Just any practice is a lot faster than good practice, than getting it right."

McDonnell added that "there is a tendency for researchers and academics to snicker and sneer at policymakers and practitioners if they don't get it right, or at public opinion that is not informed." The researcher's job, however, "is to enlighten public opinion," although often the public is more enlightened than researchers.

It could be ironic, she said, for accountability to be used to support the voucher movement, as in Florida, and she predicted that voucher proponents increasingly will make the issues of accountability and comparability of schools support their

arguments. The comparability issue will never be solved, Rosenberg said, but it also will never go away. Therefore, it is crucial to get comparisons right in order to make accountability fair. "So much of comparison reporting stops at the indicators of race and socio-economic status instead of using the data to get into deeper issues—these are taken as the end point rather than a beginning."

As to the retention issue, Rosenberg admitted that the review of research is "unassailable," but the AFT has strongly urged an end to social promotion because "all it takes is for one encounter by a citizen with a student who is not skilled to set off a problem." The AFT "has been trying for a long time to get school systems to pay attention to kids who are struggling, to give them extra help, but that does not happen so long as you are allowed to do social promotion," she said. Rosenberg noted that in the past few years, urban districts have paid greater attention to investment in summer school than in the past 12-to 15 years "when there wasn't a blessed dime for summer school." Retention may be wrong, she said, but so is social promotion, and "unless the whole issue is addressed, we will have a pendulum swing and do what's wrong, which is use both social promotion and retention."

Shepard argued, however, that policymaking is not always rational, and if summer school becomes a sham for learning, "how do I know that five years down the road the reaction will be even greater against an investment that didn't work?" she asked. There are no sure answers, replied Rosenberg, and for students who are still not getting the help they need, "you have to keep on pushing and doing it loudly." McDonnell suggested that the discourse needs to change from talking about retention to emphasizing that "we want a certain skill level by all students, which is something very different."

Validity of Testing and Assessment of All Children

Much of state and national legislation focuses on the challenge of giving all children opportunities to achieve at their highest level, Jamal Abedi of CRESST/UCLA said in opening the third panel discussion. To increase student opportunities, many policymakers have turned to accountability systems, but the issue is how valid is testing and assessment of all children.

That concern has changed somewhat the focus for the Office for Civil Rights (OCR) in the U.S. Department of Education, according to its director, Assistant Secretary Norma Cantu. "We are hearing real complaints on how assessment is used

in graduation and promotion," and it is obvious that "the new civil rights of the 1990s is higher expectations, that the civil rights community is starting to demand high quality in assessments."

Unfortunately, "the law does not hold any magic answers to the questions of the day," Arthur Coleman, Office for Civil Rights deputy assistant secretary, told the conference participants. He cited, for example, exactly opposite rulings from two judges at the same U.S. circuit court regarding the same issues in high-stakes testing. Even though numerical disparities frequently help to make a case, numbers alone are not sufficient. Nonetheless, good data should be collected to guide appropriate decision making affecting student rights.

Coleman countered what he called a "big myth"—the belief that high standards for students and equity are incompatible. "There is no inherent tension between federal civil rights laws and good educational practice," he said, but policymakers need to be aware of certain questions the courts will ask. Fundamentally, courts will always want to know the purposes of the test in question, which means, ultimately, that "education principles drive the legal conclusions."

As they insist that high stakes "be well-done stakes," courts will ask:

- **What are the educational justifications for the test?** Courts have affirmed as legitimate such reasons as improving the quality of schools, ensuring that graduates are competitive, establishing qualitative achievement standards, and otherwise ensuring that a diploma is worth something, but "the courts will want to know how these justifications are being implemented in policy and practice.

- **What is the history of discrimination and its effects?** The courts will want to know whether the potential for continued discrimination is inherent in the assessment policies. Equal opportunity to achieve the high standards must be evident, as well as how educators will address the needs of different students from different backgrounds.

- **Are the requirements new?** The courts will want to know how the requirements change from prior practice, the degree of change, and whether adequate notice has been given to parents and students.

- **What is the time between policy and consequences?** No one could seriously contend that academic requirements could never be changed during a student's 12 years in school, Coleman said, but neither could high-stakes test requirements be constitutionally imposed a day prior to graduation.

- **How is your program administered?** Are there compensatory tutorials and other academic supports, multiple opportunities to take the test, and multiple factors that enter the high-stakes decision?

- **What is the alignment between teaching and testing?** Courts will want to know the validity of the test instrument and the alignment among curriculum, instruction, standards, and assessments in the context of high-stakes graduation exams.

OCR produced preliminary guidelines for high-stakes testing, but the reactions were anything but informed. According to Coleman, hyped media reports said the guidelines would outlaw the SAT and ACT, ban all standardized tests, establish an "ethnic exception to merit," demonize academic rigor, and "place Princeton and Podunk U on equal footing." This response convinced him that "we must do a better job of talking about the value of tests . . . but also assure that the tests are being used consistent with good psychometric and educational principles."

This challenge, added Cantu, is "an education one. Those who decide when tests will be used, and how, are not academics or lawyers. They need help from those who do understand the principles. We need to find some communication strategies that take difficult, complex jargon and simplify it so decision makers understand what is going on and learn to use tests well."

The Vermont Plan

One state working on this issue is Vermont. Officials with the Vermont Developmental Reading Assessment—a large-scale assessment administered at the end of second grade—built support by including many people in the assessment's design and by starting gradually.

It was "fortunate" that state leaders in Vermont realized the importance of early reading to the state's systemic reform efforts, according to Susan Biggam, elementary reading and language arts consultant for the Vermont Department of Education. The context for taking action, however, was one of a lot of local control and a high degree of teacher autonomy. The state's approach to early reading was one that emphasized a balance between literature and skills, "with a strong focus on attention to research," she said. In order to support Vermont's early reading goals, "it was clear that a primary-level reading assessment was needed."

A broad-based task force considered several assessment options that were framed by the following criteria: The assessment must (a) be feasible, (b) match the

standards selected to focus on, (c) address Title I requirements, and (d) meet standards of technical quality. The task force also said the assessment needed to be as authentic as possible, reflecting tasks valued in their own right; link with current research; promote instruction; and yield student information that would be useful. Fifty schools volunteered to pilot the selection, the Developmental Reading Assessment. Pilot-year data analysis and teacher survey results showed that the assessment was too long and that the scoring guides were not adequate.

Modifications were made to the assessment, such as grouping the 17 short books used into proficiency stages. Researchers at the University of Vermont developed scoring guides for comprehension. The unique aspects of the assessment, according to Biggam, include annual training of teachers to conduct the assessment, individual administration of the test, the linking of early reading to standards, and the process of finding the highest level of proficiency. The assessment is designed to both provide accountability information and influence instruction. Also, it is continually shaped by research. Analysis of data, for example, showed the need to increase the expectations for oral reading accuracy and the need for a fluency guideline at the highest proficiency levels.

The developers learned that the consistency of scoring can improve over time. Biggam reported that the teachers' scoring and that of experts had 76% agreement in the first year of the assessment, increasing to 86% in 1999. In hindsight, she said, the feasibility studies helped by buying time to create a good assessment, and the standards helped to clarify the purposes. Teachers also welcomed the emphasis on improving the consistency of the standards.

A major challenge, Biggam said, is to balance the need for stability of data with continuing improvements of the assessment, noting that the developers had to add a second baseline year because of the new fluency guidelines. She also said that "there is a fine line between state control and technical assistance" in developing a large-scale assessment system.

Student Mobility and Assessments

Urban school districts often contend that high rates of student mobility unfairly affect school and district performance on assessments. Student mobility could mean students who make non-promotional or non-scheduled transfers from one school to another; or, at the school level, student mobility could represent all students who transfer out or drop out of school, a distinction that is difficult to identify. No matter

how student mobility is defined, the impact is a disruption in student learning, according to Russell Rumberger of the University of California, Santa Barbara.

The research literature on student mobility is not large, he said, but he cited two related national studies based on the same large-scale national data sets and a quantitative and qualitative study in California. The data show that there is a lot of student mobility, he said. NAEP data indicate mobility is more widespread at the elementary level than at the secondary level, and that the rate varies among students. At the fourth grade, 41% of Latino students and 45% of African American students had changed schools. In Texas, 68% of fifth graders in 1995-96 changed schools at least once in a four-year period; in Chicago, 53% had changed in a four-year period.

Using NELS-88 data, California researchers learned that most students make unscheduled school changes sometime in Grades K-12; only 40% make regular promotional changes.

Rumberger discounted the belief that most student mobility is family motivated. In Chicago, for example, only 60% of the transfers involved a change of residence. In Los Angeles, he said, the school district uses "opportunity transfers," a practice which he renamed as giving schools "an opportunity to get rid of" the students. Even if all schools enrolled the same types of students, the turnover rates would not be the same, "so we think it is what schools do that contributes to mobility."

The consequences of high mobility for both students and schools are significant. Rumberger said that high school students who have changed schools even one time are half as likely to stay until graduation. Students obtain higher test scores for each year they stay at the same school; and in schools with high mobility rates, the achievement scores of both students who transfer and those who stay are lower than in schools with lower mobility rates, controlling for other factors.

The implications for assessment systems are obvious, Rumberger noted. Mobile students may be more difficult to test because often there is a gap between the time when a student exits one school and enters another. A student could be out of the system for weeks, he said. Moreover, mobility makes it difficult to assess schools. Mobility is so disruptive to student learning that many assessment systems exclude students who have transferred because there is an assumption that schools can do nothing about the problem. Rumberger disagreed, saying, "unless schools are given

incentives to keep existing students, high levels of mobility will remain." He believes schools "should be held accountable for all students who walk in the door." They could create cohort assessments of students over four-year blocks. Also, all high schools should be judged on their graduation rates. Without that, schools have incentives to get rid of some students. Rumberger noted that in California in 1998, alternative enrollments grew at four times the overall enrollment increase.

Assessments and Student Language Background

Math assessments are not language neutral, Jamal Abedi, CRESST/UCLA, emphasized in reporting research on the impact of language background on NAEP math scores. Focusing on the effect of accommodations, he said, "there is no such thing as an accommodation that is good for all. Its effectiveness depends on the background of the student." The research is difficult to carry out because of the lack of control groups among the non-LEP student population. As part of their studies, Abedi and his colleagues made linguistic modifications of NAEP test items (e.g., shorter sentences, simpler rules, and elimination of unnecessary expository material). For example, the word "When" was substituted for the phrase "At which of the following times." Students found the assessments with linguistic modifications easier, and their speed in answering questions was faster.

LEP students' scores were higher on all types of accommodation except when they were only given a glossary. Modified English, extra time, and use of a glossary *plus* extra time helped them. The best predictor of math scores, according to Abedi, was the length of time the student had lived in the United States. Other predictors were how far the student expected to go in school, how good the student was at math, and how many times the student had changed schools.

Impact of Accountability Systems on Instruction and Outcomes

Beyond the ubiquitous complaint that accountability systems lead to an inordinate amount of time spent on test preparation—the "teaching to the test" syndrome—how do tests influence instruction and outcomes? The news from an expert panel on the subject was both good and bad.

The process in the Pittsburgh public schools is a case in point. Lauren Resnick of CRESST/University of Pittsburgh described the school system as "right in its core thinking on math." An enthusiastic supporter of standards and an original member of New Standards, Resnick said that Pittsburgh's core curriculum frameworks were

heavily influenced by the National Council of Teachers of Mathematics (NCTM) and New Standards. Pittsburgh administered the New Standards math reference exam to all fourth and eighth graders in 1996, and two years later adopted the New Standards as the standard for which schools would be responsible, "giving schools a strong message that this assessment was fully aligned with their standards," she said.

In 1996, the school district also adopted the National Science Foundation-supported *Everyday Math* program, which also was aligned to NCTM standards, and received a National Science Foundation grant for professional development in the schools. All the elements then were in place, according to Resnick, including standards, curriculum, professional development, and assessment.

Student achievement did not improve much "until the whole package was implemented," Resnick said. Matching schools on demographics, researchers found that in schools where there was weak implementation of all of the components, African American and White students' scores were equally low. In schools with a strong implementation, the scores of both African American and White students were high. African American students in strong-implementation schools outperformed White students in weak schools, she said, "practically wiping out the racial gap."

Even though the district was aware of the data, it did not follow through on accountability, according to Resnick. "If I were a superintendent looking at the data and seeing that we can actually raise achievement and go a long way toward closing the achievement gap between minorities and Whites, I would be saying we should be stronger in our accountability demands," she said. However, the district's accountability system was weak. It did not call upon teachers and administrators to use what had been shown to work, so the effects were scattered "and all of the standards-based activity has much less value than it otherwise could have." Interviews with area superintendents revealed that they did not systematically implement the reforms.

"We have evidence from a lot of places that poor kids can learn at high levels where there is systematic implementation," Resnick said. "Do we not have an obligation to use what has been proven right?"

In the discussion following the panel, Resnick explained that a "strong" school was one in which all teachers in Grades 2-4 had participated in focused professional

development and had implemented standards-based instruction for two years. The "weak" schools were those where hardly any teachers had done that. The policy goal, she said, is high standards for all, not closing the gap. The accountability that might change performance probably is not any formal counting system, such as who is going to fail or be promoted, Resnick said. Rather, "it may be something as simple as the supervisor of principals asking them how many of their teachers have participated in professional development, how many are teaching the curriculum, and for those who are not doing it, why. My prediction is that within a year, 85% to 90% of students would be climbing up in achievement."

The Long Look From Kentucky

Kentucky's accountability system, started in 1989, is well evolved and typical of what is being discussed in other states. As a result, the research by Brian Stecher (CRESST/RAND) and others at RAND and the University of Colorado at Boulder is applicable to accountability systems elsewhere. At the time the studies began, the Kentucky assessment followed the pattern of the National Assessment of Educational Progress (NAEP), testing one subject at different grades, either through open-response questions, or portfolios, or a combination. The research studies included two rounds of interviews with teachers of math and of writing.

The assessment carried high stakes in that schools received bonuses for doing better than expected or, at the other extreme, were placed on a crisis list and put under strict supervision by the state. The testing, according to Stecher, "has led to a lot of changes in instruction that seemed to be positive, but teachers are focusing quite narrowly on the tests, particularly in the scoring rubrics." The assessment also places considerable burden on teachers in the grades affected, especially those whose students must prepare portfolios.

Stecher also said that the effects are very uneven across the grades. "Teachers are extremely nearsighted, focusing on their own grade in terms of professional development, allocation of instructional time across subjects, and the frequency and type of test preparation activities." Researchers found, for example, that the mean hours per subject in a typical week in self-contained classrooms differed dramatically, depending on which subjects were being tested in their grade level. "To me, this does not make for an effective education program, particularly when students will be tested again in a subsequent grade on the other subjects," he said. "There is no evidence that this is bad, but intuitively it seems to be so."

Generally, as in Kentucky, a standards-based system sets performance standards, which are to be the impetus for change throughout the system, including professional development and classroom practices. Testing programs are intended to tell what is being achieved with the hope that student performance reflects student learning. Adding high stakes to this process heightens public and schoolwide discourse on what happens on the test. But as Kentucky illustrates, "tests become more salient than the standards, and classroom behaviors are more about changing scores than changing what students know and can do," Stecher said. This creates "perverse effects" in a standards-based accountability system.

Stecher recommended several ways to limit the negative aspects of accountability. Replace milepost testing with testing in all grades. Broaden the range of assessments in a number of ways so that teachers focus on domains of interest rather than test format (e.g., use multiple assessment formats and/or matrix sampling of content). Revise the test forms annually. Reduce the stakes. Set reasonable targets (Are world-class standards realistic?). Emphasize standards rather than assessments through professional development and standards-based curriculum materials. Create an external-testing audit, as does NAEP.

Although the technology exists to produce an assessment system that gives results, the tests are not "tweaking" instruction, Dan Koretz, of CRESST/RAND/Boston College, said in a presentation that built on the findings in Stecher's work. Some states are making the assumption that tests are sufficient to drive instructional improvement, but instead, teachers are shifting time and educational resources back and forth, depending on the testing schedule.

Comparing the gains on the Kentucky assessment with gains on NAEP, Koretz said there was an overlap between the two that reveals the gains on the state test "are outright bogus and a reflection of coaching." There is long-term evidence of predictable gains on tests, with the only question being whether it takes two or four years to happen, and he cautioned against anticipating evidence on gains that can be generalized. Citing an experiment of a decade ago, Koretz said third graders were administered a test no longer used by the district, and as with most new tests, students tested low—at first. Test results went up every year after that.

Although the Kentucky Instructional Results Information System (KIRIS) and the NAEP are quite different, the former was built on the frameworks of the latter. The evidence on KIRIS indicates the gains were "exaggerated," according to Koretz.

In fourth-grade math, for example, they were identical to the national average on NAEP, resembling gains made everywhere else. KIRIS showed very high improvement, but ACT results remained stable. The problem is that schools in Kentucky were told to make unrealistically large gains—two standard deviations over a 20-year period, "and so they did," he said.

The assumption of such accountability systems is that they can make the variance in performance disappear, but the faulty thinking shows up in comparisons revealed in the data from the Third International Mathematics and Science Study (TIMSS). "We stratify kids more than in Japan and Korea," Koretz explained, so that variances show up within classrooms, but are not predictable between classrooms. The bottom line is that "there is not a shred of evidence that we can reduce the variance more than 20%."

Accepting that there will always be big variances, Koretz said tests should be designed for specific uses and be monitored for inflation of gains. Discount or ignore the initial gains—"Tell the public to not get excited about the first two years of results," he said. Policies should set reasonable expectations for the rate of improvement and for variance in outcomes. "It is never going to be true that all kids will perform the same, not even true that all kids from the same backgrounds will have the same performance," he said. "There will always be some kids whose career interests do not include calculus."

In the discussion following the panel presentations, conference participant Cheryl Tibbals, director of KIRIS for the first three years of the Kentucky reform, recommended the following:

> Researchers could greatly assist educators in the field by taking into consideration the political and educational context in which states are working. While educators may wish an open window of time to turn the performance of students around statewide, they frequently are not given that amount of time by the public or legislature. While Kentucky educators may have wanted even more time to get all students to higher levels of proficiency, using 20 years as the time frame was better than the 2 years which the Legislature would probably have preferred.

She and Koretz disagreed on reform goals, with Tibbals defending the expectation that all children can learn to high standards and Koretz saying that, while the lack of opportunity in poor schools is "deplorable," children do differ in their ability levels.

Accountability in Higher Education

There has been a significant increase in the demand for accountability from higher education, but it could be a case of déjà vu, according to Richard Shavelson of CRESST/Stanford University. The flurry of interest in the late 1980s, spurred by the U.S. Department of Education and accrediting agencies, died out, but the issue is back again and more complex than ever.

Applying for-profit sector principles to the non-profit world, as proponents of accountability suggest, is problematic, Shavelson said. The bottom line in business output measures is closely aligned with valued outcomes such as revenue. In the non-profit sector, goals are multiple, not quite so clear, and not readily measurable. In higher education in particular, it is impossible to reduce the impact of college to a single goal because campuses are interested in citizenship, social interaction, and respect for others, as well as academics. "What worries me even more than this context is that we are not discussing it at all," Shavelson said. "We move forward and pick achievement indicators, and cost and efficiency—it's gotta be cheap and fast—always are part of the picture." These achievement measures become proxies for what is wanted from education, and the proxies become the outcomes. "We are dealing with a very delicate system, and there is a lot of room for harm," he said.

Higher education uses a variety of models for accountability. The internal audit, recommended by Patricia Graham and others many years ago, allows each institution to decide on its valued outcomes, then collects data on academic outcomes by department. It may be regarded with suspicion because the university is producing its own data. An external audit looks at such things as the number of publications tied to budgets—"a campus would have to grow a lot of journals to stay in business," said Shavelson. Another method is to monitor indicators that are changeable.

Taking cues from the experiences with accountability in the K-12 system, Shavelson said that concerns about the impact of accountability should be shared "because of the possibility of mischief. It is possible that the accountability results in distortions of the curriculum so that it becomes a mile wide and an inch deep, teachers teaching to the test, schools cheat, and there's a drift upward in average test scores."

Possible design principles for higher education, according to Shavelson, might include an expansion of the notion of achievement, alignment of formative and

summative assessments (the former to help in classroom improvement, the latter to be used for external reporting), a recognition of tremendous variability in institutions, and clarity on differentiating purposes.

His research is trying to map out cognitive domains to expand the concept of achievement, such as defining declarative, procedural, and strategic knowledge. Formative and summative evaluations must be aligned, and accountability systems must honor variability. Shavelson said that in the Los Angeles area alone, "we see radically different learning environments." If these principles are used, then the good news is that "accountability will move forward and can improve teaching and learning in college classrooms because faculty will know where they need improvement." The bad news, however, "is that we can make a mess of this and wind up reducing outcomes to the lowest common denominator."

In summarizing the themes of the panel, Joan Herman of CRESST/UCLA said assuredly that there are ways to guard against the potential misuse and distortions of assessments. "The answer lies not in trying to micromanage from afar what schools should be doing," she noted. "Assessments in accountability systems provide only general indicators of how things are going. They are never going to tell you what you need to know to truly understand students' learning and how to improve it." To get that, "we need to give schools and teachers tools to do it for themselves; but the more pressure and high stakes placed on tests, the more we are going to drive out the talent we need in schools if we are really going to make a difference for kids."

Small-Group Discussions

Conference participants discussed the politics and the particulars of ongoing research projects in a series of concurrent sessions held each day of the conference.

Technology Use

The IMMEX (Interactive Multi-Media Exercises) project of the UCLA School of Medicine now has four years of experience in providing K-12 teachers with professional development in technology usage in their classrooms. Ron Stevens, the project director, can say with certainty that professional development in technology "does not guarantee that technology will be introduced in the classroom." Instead, there are sequential stages in professional development and technology use.

Originally used with medical students, IMMEX evolved into a K-12 activity with the goal of encouraging teachers to create technology-based curriculum, integrate technology into classroom practice, shape students' metacognitive skills, and become researchers. The project offers training institutes where teams of teachers work on software for their classrooms based on problem solving. The software also tracks and records what students do as they go through problem solving, diagramming the processes students use (e.g., scattered or orderly searches for information).

The professional development consists of month-long training institutes, which started with 20 teachers and are now up to 100 teachers per workshop. Most of the teachers who attend have not used technology in their classrooms. They come primarily from the Los Angeles Unified School District, but the project has extended to partnerships in Orange County and Pasadena as well as to teachers from outside of California.

"We thought that if we give the workshops and have good outcomes, technology would be used in the classroom," Stevens said. However, few student performances were recorded in the first two years. By 1998-99, the project recorded 12,500 completed student performances with 150 different software problem sets. Of the 300 teachers trained at that time, 55 were active users of IMMEX. The most popular uses were in chemistry, biology, math, and genetics. In the summer of 1999, the project held a special workshop where teachers conducted data analysis of the level of difficulty, the performance differences between classes, and what types of students were using the software. Funded by a grant from the National Science Foundation, the project spends $6,600 per teacher for the program, but as more students run the problem sets, the costs will come down significantly, according to Stevens.

Over the years the program has increased variety in its professional development to take care of all needs, Stevens said, but the differences among teachers show up after the institutes. Some teachers from previous institutes are now mentor teachers at subsequent institutes; others are galvanizing interest in their own district, giving workshops for all teachers. Some teachers are still reticent about asking for help and "don't understand the dynamics yet."

Using IMMEX teacher surveys and interviews, Gregory Chung of CRESST/UCLA found that using technology in instruction required teachers to

have a lot of experience in conceptualizing where problem solving can best be used, a more fundamental problem than the lack of computer skills. IMMEX provides considerable support for teachers, including stipends for the training, "but implementation is still low," Chung reported. High users "buy the whole package," and "look for opportunities to use technology as a tool to help them achieve learning goals." Moreover, these are not time-clock teachers, he said, and they are willing to spend time learning and using IMMEX resources.

Chung recommended that schools build on the interests of those teachers who implement IMMEX, predicting that the technical barriers eventually will disappear. Teachers, however, "will still have to go through a developmental process of figuring out what they want to present in content and how to structure it."

One of the high IMMEX users, Paula Dallas of Pacific Palisades High School, said the training helped her overcome a lack of technology know-how and find assessment tools that are very effective in her highly diverse classrooms. The software allows students to have equal opportunities to show what they have learned. Instead of comparing student A to student B, she said, "you are comparing their original data to their second try at it. You can see a kid's depth of understanding, what data they used, and their search path maps. Also, kids have their own opportunities to assess themselves."

Dallas uses the IMMEX problems on a regular basis, especially in the study of genetics. The project collects data and sends back an analysis of search path maps, allowing her to see whether her students have narrowed the focus of searches to solve problems. Most of all, she said, "my students have experienced an opportunity to problem solve that many students haven't had. They are developing into critical thinkers."

Studying the effects of using the Web in the classroom, Davina Klein of CRESST/UCLA reported that teachers have moved from being experts to being partners with students. With more than 800 million pages of information now available on the Web, students need to be able to navigate through large information spaces to find specific information, like searching for a needle in a haystack.

Using a closed Web system, researchers asked students to do research, tracking everything they did, including what they bookmarked. They compared students' database searching to what expert searchers do, use of key words, and the efficiency of their searches. Basically, the students, in Grades 5 through 12, were not very good

searchers. If they were good searchers but poor navigators, they could not find the information. The lesson, Klein said, is that if technology is going to be used in the classroom, the expected benefits must be defined, measures sensitive to those outcomes must be created, and there must be assurance that "everything you expect is happening before you get to the level of assessment."

Lessons From Implementing Assessment Systems in Large Urban Schools

Lesson 1: Those outside of the classroom consistently underestimate the size and scope of assessment programs. This was the opening gun from Robert Collins, assistant superintendent of the Los Angeles Unified School District, which serves more than 700,000 students. July is happy month in the district, he said, because it is the only month when the district is not testing. Administrators tend to look at assessment from a global viewpoint, academics see assessment as driving the whole train, "but at the school level, educators are being beaten on by many different assessments, sometimes all in the same two to six weeks."

A second lesson from Collins, a former high school principal in the district, is that assessment is consistently depersonalized. "In the media and central office, we discuss the issue in terms of median and means, gross numbers, percentiles, and all those other lovely things," he said, "but at the school, it is about 400 students who aren't meeting grade-level standards." In the district as a whole, there are 170,000 students currently not on grade level and at risk of not being promoted. When looking at all the facts and figures, district administrators and others miss the individual student.

This is an error, according to Collins, because it leads to a failure to use the assessments to provide the right instruction, purchase resources, and develop strategies that can be effective. "We also tend to forget who these students are—mostly minorities and limited English [proficient] learners—as does the media," he said. "Individual schools are hit with the results, and then try to explain to the community or to the media what test results really mean, but they might as well be talking to a wall."

Collins also said the enormous faith put in assessment instruments—"and almost total reliance by the media"—leads to a ranking of schools, holding people accountable based, in some cases, on a single instrument. "The use of multiple measures is the only survival factor in this whole equation, such as teacher judgment and teacher marks combined with performance assessments," he noted.

He also said that there is a tendency to overestimate the significance of testing. It no longer is used to tailor an education program to assist both the teacher and the student. Rather, testing "is now used to rank schools, fire principals, hold teachers accountable, and hold students accountable in high-risk endeavors."

Collins' other lessons: Parents too often are not part of the assessment process; assessment and assessment processes are consistently underfunded ("performance-based assessment can't be run on a shoestring budget, it must have a chunk of the budget"); assessment programs must provide teachers and principals with the appropriate amount of professional development time; and assessments for limited English proficient students and students with disabilities should be developed as part of the overall program, not in isolation.

"The thing we do worst of all is make assumptions about what is occurring in schools at any given time," Collins said. For example, "we assume all teachers come to us with a belief system that all children can learn, but many come with the idea that 40% of the children will never learn." Teachers cannot be trained effectively on standards and rubrics if they don't begin with the idea that all children can learn, he commented.

It takes sixty-two 18-wheel trucks just to deliver the state's Stanford-9 test to Los Angeles, where schools already are dealing with a multitude of other types of exams, Collins said. "We have been working for three years on the assessment system, and still feel we are behind. Don't ever believe it is a simple, uncomplicated process," he said.

Working in Los Angeles and Chicago to build standards-based assessment systems, David Niemi of CRESST/UCLA agreed with Collins that just setting standards and deciding how to test them would not suddenly help all students learn. The public conception that results will be immediate has to be confronted all the time, as does explaining how every student cannot be above average in a norm-referenced assessment.

Another issue is test preparation, because it often doesn't focus on content understanding but rather on how to deal with a certain kind of test. This focus on the test at the expense of what it is supposed to measure is "unfortunate," Niemi said. Also, states are continually changing their assessments after school districts and teachers have adjusted to the standards of the current ones.

To get curriculum aligned with standards, the standards must be measurable activities that can be done in the classroom, Niemi said. Some kinds of standards, for example, cannot be assessed very well by multiple-choice questions. The standards in California call for students to be able to write a clear and simple composition, but the state is using a multiple-choice format to assess that ability.

To mitigate such problems, CRESST is using a strategy of looking at subject-area research as the basis for figuring out what standards mean to teachers, how to assess standards, and how to score and interpret results. To answer what it means to know a subject area, the researchers studied what mathematicians and scientists know, which produced some commonalties. "Advanced subject knowledge usually is characterized by a high degree of organization and ways of connecting things together," according to Niemi. "People know when and how to use their advanced knowledge, they know what it is for," he said. When working with teachers, however, Niemi has found that sometimes not all of them have a high level of knowledge about the subject area that is being assessed. To get that knowledge, teachers are learning from the researchers how to analyze the standards to be assessed.

California's strategy has been to take norm-referenced items and match them to the state's standards to produce a modified, off-the-shelf assessment. The CRESST approach developed an assessment that is sensitive to instruction. "If you don't teach the concept, students won't do well on the tests," Niemi said. It has been a three-year process of working with teachers, starting with the analysis of standards. Now the process is at a point where there is a large cadre of teachers who can do the analysis. The project is moving into regional clusters and will affect all schools in the clusters.

In the scoring aspect, teachers look at a variety of student papers and decide what makes good work—what are the kinds of principles students should be demonstrating in their work. CRESST is training a representative from each school on how to do the scoring; those teachers will train other teachers, and then all teachers will score student work together in groups—for example, all third-grade teachers will work together.

Scoring is done at the school site; in the first year, 20% of the papers were collected and re-scored centrally, with better than 50% agreement. Eventually, the

focus will be on the relatively small number of teachers who need to improve their scoring skills.

Explaining even further the role of teachers in the development of this assessment system, Gina Koency of CRESST/UCLA said the pilot tests asked for teacher feedback on such questions as the clarity of the rubric and the difficulty of items. "These kinds of activities appeared to make a difference," she said. "Teachers who had used performance assessment infrequently or not at all realized the kinds of assessments of student knowledge they could make after using the pilot tests." A consistent message was that they were going to use such assessments on a more regular basis. Even in classrooms where students struggled with the assessments, teachers were less likely to say they were too difficult. Rather, teachers believe now that students ought to be meeting the standards reflected in the performance assessment.

Although teachers wanted to participate in scoring, at first they just ranked student papers from high to low, without a rubric. Then a group of teachers designed short units of instruction around standards and assessments, randomly assigning them to students. Koency noted that the results of performance assessments were promising. Teachers were surprised at how little instruction was needed to get students to do well if the instruction was clearly focused.

The study found that information about testing is not always disseminated and conveyed to teachers as intended and needs to be decentralized.

The project gradually has turned over much of the design work on performance assessments to teachers who have been involved in the process. "Once they learned the concepts, they became more focused and creative on assessments," Koency said. Small groups, fewer than five at the elementary level, provide the most effective learning situations; focused, intensive workshops conducted over five consecutive days are better than workshops spread over a year. Teachers enjoyed the opportunity to collaborate, she reported, and wanted to do more of it within their schools and grade-level teams.

"Where teachers meet on a regular basis, analyze data, implement an assessment, and analyze again, there are significant improvements in student achievement," Koency said.

Reporting School and District Progress

Schools are characteristic of poor information design, according to Gregory Leazer, professor in the Department of Information Studies at UCLA's Graduate School of Education and Information Studies. Poorly designed information systems occur when people have information, knowledge, or understanding but these are not communicated through the system. "The assumption here is that schools have lots of knowledge about local performance," he said, "but they are not doing a real great job of communicating that out to policymakers, the community, or parents."

One of the most important elements in information systems is the "user interface," when the knowledge connects to a user, Leazer explained. A tragic example of a weak user interface is the explosion of the Challenger space shuttle. Engineers actually predicted disaster if the spaceship were launched on a cold day, but as their reasoning went up through the system, the user interface failed to be persuasive. The engineers' arguments lacked explaining charts, used visual clutter that obscured the data at hand, failed to make clear the cause and effect, and had poor ordering of data, Leazer said.

A good information system for schools would not make any of these mistakes, Leazer said. The bottom line for school information systems would be to convey an understanding of the dynamics of variables and a deeper understanding about schools and their performance. School report cards need to inform the public and government bodies, to inform educators and education researchers on necessary improvements, and to identify schools for rewards or punishment. Yet, he added, report cards don't provide precise interpretation.

About 36 states now produce annual report cards. Twenty-six of them post the report cards on the Web, but half of that group also require the cards to be sent home. Test scores are the most frequently presented data, but the report cards also include data on dropouts, absenteeism, and Advanced Placement enrollment. A survey of parents revealed, however, that they wanted most to know about school safety and teacher quality. Student scores were far down on the list. The least important information to parents was class size, graduation and dropout rates, and student demographics. The cards allow people to make comparisons among schools through clusters, national/state averages, individually identified goals, or what the school did in previous years. One problem Leazer said he encountered in working on report cards in Los Angeles is that the information differed according to the

source, with the state, district and local schools "clearly having different methods for generating data." Only 12 states assign grades to schools, although 7 others identified low-performing schools.

Leazer's analysis of the formats for the report cards found a lot of tabular data that was poorly communicated, "with a lot of time spent on a dashboard or cockpit format." In helping Los Angeles design a report card, the decision was made to include only information under the control of a school or over which the school has some leverage. The Los Angeles report card is a collaborative effort of the Los Angeles Office of the Mayor, UCLA, and the Los Angeles Unified School District, using a single-page graphic presentation of a whole school.

Research is needed on the most appropriate use of report cards. At Fremont High School in the Los Angeles district, for example, students went out on strike because of the number of substitute teachers being used, but the report card did not reflect such problems at the school.

Whereas the report cards Leazer is developing are intended to inform others about a school, the Quality School Portfolio, also a project at CRESST, is meant to inform a school about itself, explained Derek Mitchell of CRESST/UCLA. QSP "sits" at the school site, absorbing data from the outside, but a school also has the flexibility to add data to it—for example, its own surveys based on questions important to it. QSP makes school-based action plans more achievable and can send out information to all audiences, such as teachers in certain subject areas, students taking the teachers' classes, and their families. The initial schools were chosen to participate in the project because they had someone to take charge on site, were involved in some kind of reform, and wanted to use data to do things differently.

QSP includes a data manager and a resource kit. The data manager covers basic descriptive statistics that become more sophisticated as the data generate further questions. The project gathers this information for the school, although this has been more of a challenge than anticipated. In the summer of 1999, the project began QSP training institutes in Chicago and Los Angeles on the use of the data and on self-evaluation. The project will follow up with site visits. The three-day training for school teams included use of the data manager and how to provide a context for the process using a school's own data sets. All schools except one decided to gather non-cognitive data such as student surveys or the demographics of technology use at the school.

One of the weaknesses of the project in Chicago is that the school district is not geared to providing school-specific data. For example, the district does not keep data on how long students have been at any particular school, so schools have to go through their own rosters to figure that out, a problem that also was true for many schools in Los Angeles.

Schools rely on different staffing arrangements for QSP. The "data evangelist" model uses one person at each school "whose job it is to drive the questions and keep the data set on a single computer," Mitchell said, a situation that can either create a data Napoleon or overwhelm one person. The collaborative model, used by most schools, calls for a school site team to set the agenda and drive the questions. Sometimes this means there are too many cooks in the process, leading to schizophrenia in the data because of different lines of inquiry. The distributive model creates multiple QSPs at each school, which may or may not have the same data. For example, the reading and math coordinators may have their own data sets, as well as the data sets of their departments.

In Los Angeles the schools are using the data manager to disaggregate SAT-9 scores, probe the contribution of afterschool programs, or analyze feeder clusters, as examples. Users often want QSP to make inferences from the data automatically, but human judgment is still important, said Mitchell. "QSP can only tell them that these students are strong and these are weak" in a subject, for example, but it is up to the schools to determine whether the differences are related to teaching.

Mitchell has learned that school personnel are not very interested in talking about standards and have little interest in asking questions about the alignment of their curriculum to standards. There also is very little interest in professional development in the area of evaluation. QSP is improving over time with new reporting functions, an ability to sort by multiple variables, and the ability to edit reports.

Using Multiple Measures

The issue of using multiple measures in assessment is very complex, interesting, and important enough to attract research money, said Dale Carlson, somewhat facetiously, as he opened a session on the topic.

Formerly with the California Department of Education, Carlson said that multiple measures are all around us and existed long before current education law.

But they engender a number of questions and concerns, such as making a distinction between multiple measures and the combining process. "What makes measures multiple?" Carlson asked. Multiple instruments? Multiple formats? Giving students multiple opportunities?

Carlson is looking at multiple measures in two contexts: the levels of analysis (students, schools, and states) and the purposes of the assessment (accountability, school improvement, and improving instruction). The process begins at the student level with achievement data that aggregate up to the school level, and then are combined at the school level. A composite multiple measures system uses three realms of thinking and research, Carlson said: statistical, empirical, and logical/rational. If evidence is put together from these three types of knowing, then we should be able to develop criteria and specific procedures that will allow us to combine the numbers in ways that make sense according to both the purpose or use of the combined result and the characteristics of the data. Carlson said that both empirical research and hard thinking would be necessary before we really know how to do this.

Alan Sheinker of the Wyoming State Department of Education provided examples of how information can be combined in a multiple measures system. The conjunctive model uses separate assessments by subject area, and a student must pass them all. The disjunctive method requires students to pass only parts of the assessments by weighting the performance standards. The compensatory model allows one test to compensate for other tests so a student could perform lower on one instrument within the parameters set.

When talking about multiple measures with teachers in Wyoming, Sheinker said he asks them to map the standards and evaluate their coverage rather than focus on the test. He warns against a single assessment measuring one dimension and noted that some districts are using mixed models, combining both the compensatory and the disjunctive approaches. There are significant issues in determining adequate yearly progress, as required by Title I, in small schools. Sixty percent of the schools in Wyoming have fewer than 100 students in each grade, so a low school one year may be a high one the next "because a bright kid moved in." The models must yield quality results for decisions, then look at the variables that impact those decisions, he said.

Using fourth-grade achievement scores in reading, math, and writing, a CRESST project compared the three models, looking for descriptive and relational differences and a stability index, according to Barry Gribbons of CRESST/UCLA. Most schools were in the 25-50 range, but under disjunctive and conjunctive models, "the distribution comes crashing down," he said, with conjunctive the most stringent (distribution was at the bottom). The compensatory was "in between." The rank ordering of schools differed according to the model used. All correlations were much lower on a stability index.

In sum, Gribbons said, "I can't offer a lot of guidance on which method is most stable." He did note that at the high school level, many more quantitative indicators are available, such as AP classes, grade point averages, dropout rates, and attendance rates. The differences between middle school and high school data are small, he said, but they differ greatly from elementary school data.

Measuring School Performance: The California Academic Performance Index

Approved by the California legislature in April 1999 as an accountability measure, the Academic Performance Index (API) is on a fast track despite its high stakes, according to William Padia, director of the Office of Policy and Evaluation for the California Department of Education. The legislation sets up an advisory committee that guides the state superintendent and state board as well as a larger, 42-member group consisting of representatives of all stakeholders; within this structure there is an API advisory subgroup.

A second part of the legislation required the development of the API—one number that ranks schools. A third aspect—the Governor's Performance Awards—depends on growth in achievement and provides direct financial rewards to schools and teachers. A fourth part covers intervention in underperforming schools. On a voluntary basis, schools would get a planning grant, then implementation grants for two years; they would get off the list of underperformers by doing well on the API. From a pool of 3,100 schools out of the 7,000 in the state, the intervention project received 1,400 volunteer requests and selected 430. The last part of the legislation calls for evaluation.

The legislation gave the state department only three months to develop the API. The agency said that was impossible and, instead, provided a framework for guiding principles, Padia said. The principles say the API is to be used to measure the performance of schools, which must show comparable improvement by all

significant subgroups. "You can't give a good school an award if there are numerically significant groups within the school that aren't doing well," Padia said. At least 60% of the API must be based on student achievement, but a number of other indicators are required as well. They include the state Standardized Testing and Reporting (STAR) system, attendance rates of students and staff, and graduation rates at the high school level (the latter two will not be available for several years because current data are not accurate).

The API is a growth system, Padia explained, because everyone starts from an individual baseline, but the state is required to do a number of rankings with these data (e.g., ranking by deciles within levels of schools). The baseline will consist of two years of STAR data and the SAT-9 data from 1999. The first ranking will be the actual value of the API, and a second ranking will be the decile ranking of similar schools. All of the information will be posted on the Internet.

As a member of the advisory subgroup that recommended the principles for the development of the API, Eva Baker, co-director of CRESST, said the whole effort focused on "trying to create a valid, useful, credible system that will serve the children of the state." The greatest difficulties facing the subgroup had to do with the legislation itself—the tight schedule "and the notion that things could be squeezed in," said Baker, adding that it was better to participate and deal with the constraints than abstain.

In addition to being technically sound, the API should emphasize student performance and not educational processes, according to the principles. "We wanted schools, parents, and the community to focus on how well children were doing," Baker said. It also needs to "strive to the greatest extent to measure content skills and competencies that can be taught at school and measure state standards." However, there will be tradeoffs, she admitted, because the goal is to create a system that will be valid in the long run and more accurate as it goes along, even though there are some short-term tradeoffs that will be made.

The API should be inclusive, with mechanisms to deal with the issue of waivers and no-shows. The legislation mandated 5% growth, "but that doesn't do anything about raising the bottom line in schools that are not performing," Baker said. The issue is not one of closing the gap. All models did not change the fact that higher performing schools will have an easier time, meaning "there is no solution to this problem other than to get lower performing schools to accelerate." So, the

advisory subgroup recommended that the model look at 5% growth as a target. The API must be accompanied by comprehensive information and not be seen as a supplement for local accountability measures that districts feel are important, the principles also state.

"If we applied these principles to any accountability system out there now," Baker said, "almost every one would fall short. There is no paradigm to use; it comes down to a value choice among the principles."

The technical subgroup faced three challenges, according to Padia: "how to calculate this thing, what are the errors, and what are the effects on the system." As to the calculation, the developers recommended that students be grouped by quintiles, using the 1999 STAR as the baseline, and weigh efforts for moving students from one quintile to one higher, with more weight given for moving students off the bottom quintile. There are 300 points for moving students from level 1 to 2, and gradually fewer points for higher levels. The API comes from the multiple value of the quintile times the percentage of students in the quintile.

At the elementary level, Padia said, the index is weighted more toward language arts and reading (60%) than math (40%). At the high school level, equal weight is given to the core subjects (mathematics, reading, language, history, science). The 5% growth rule caused trouble, making the developers choose between the target or base year and consider whether the target should be more for low-performing schools. They decided to make the target the same for every school, but wondered how high-performing schools can show progress on the SAT-9 because it is a basic skills test.

As to what the API will reveal about actual learning in the classroom, there was some skepticism. Using gain as the major element "makes this goal a little more palpable," Padia said.

Baker commented that "policy statements that get into legislation may be a good idea, but they take a lot of technical work. We have been trying to come up with a system that is as simple, rational, and fair as possible, but we may fail because of the lack of clarity in the policy." It will be difficult to explain to schools what to do with the information, she said, because "this has to be complicated in order to work."

Issues Surrounding College Entrance

The elimination of affirmative action in college admissions in California did not do away with higher education's commitment to diversity as a critical value, but it certainly made it more difficult to accomplish. When developing new policies, the state's three-tiered higher education system—the University of California (UC), the California State University system, and the California Community Colleges—kept to distinctions between eligibility (systemwide) and admissions (by individual campus).

Explaining the new policies, Saul Geiser, manager of research and planning for student academic services in the Office of the President at UC, said the eligibility pool had been redefined to include percentages from each of the state's high schools. The UC system captures the top 12% of high school graduates; the top one third is eligible for the Cal State campuses.

With no statewide database on high school students, the planners were forced to use the College Board data. About 46% of all California high school graduates took the SAT, and 98% of the applicants to the UC system took the SAT, "so there was a good overlap with our requirements," Geiser said. The academic index scores for all high schools were computed to get the rank order of high schools, then the number of graduates from each high school was compiled, and the eligibility pool was taken from the top end. This calculation gave the UC system a desirable racial/ethnic mix, but it would mean a decline in academic measures. The faculty became interested in the process, recommending that more weight be given to college preparatory work. The final plan reduced the eligibility pool to the top 4% of high school graduates, with rankings based on grade point averages (GPA). This doubled the African American and Latino existing eligibility pool from the lowest SAT percentiles, but the students were considered to be "at least competitive and should have a very good chance of graduating from the university," he said.

The UC system is now giving more emphasis to family income to "help attain some semblance of diversity in the admissions pool" added Tom Skewes-Cox of the Student Affairs Information and Research Office at UCLA. By considering both the GPA and the SAT, "we are looking for a student who took a strong program and got good grades," he said, but using income is helping to maintain some sense of diversity. Between 1997 and 1999, the university experienced a 40% drop in

underrepresented students who came from all over the spectrum, but particularly in the middle quintiles.

"In the face of trying to maintain diversity in higher education, it has become apparent that the traditional approaches to outreach will not be enough to ensure a diverse representation of students at the most selective University of California campuses," where the average freshman has a high school GPA above 4.1 and a combined SAT score above 1310, said Denise Quigley of CRESST/UCLA. Understanding these constraints, the UC system has partnered with a total of 70 high schools across the state to raise the competitive eligibility of their students. At UCLA, the efforts include a comprehensive and theory-based program called CBOP, the Career Based Outreach Program, which integrates service learning with teaching students how to be optimal learners.

In the first year of its implementation, reported Quigley, CBOP did not seem to have an effect on the participating ninth-grade students' achievement, course-taking behavior, or study skills. However, CBOP did have positive effects on the UCLA undergraduates who mentored the high school students. "This program and its unique service-learning design have promise," said Quigley. The next few years are crucial in understanding its impact on high school students, she added.

Systemic and Classroom Perspectives on the Alignment of Goals, Instruction, and Assessment

Bringing a viewpoint from reforms in Australia, David Clarke of the University of Melbourne began with a universal observation: "Curriculum change frequently fails to do more than change the paperwork and the rhetoric that teachers and others exchange in conversation. It does not change practice."

Assessment should follow instruction, but as many presentations at the CRESST conference had alluded to, assessment may drive instruction. That observation is frequently pejorative because, Clarke said, "if teachers, students, and parents cannot look to our system and find a model of the practices we want, where are they going to find them?"

In Australia, the pivotal time is the final two years of secondary school, when assessment is mandated. College entrance is based almost entirely on the assessment results. To study the assessment impact, researchers looked back to the seventh grade, collecting all forms of school documentation for quality math performance

including school policies, syllabi, and worksheets. They then constructed questionnaires "to unpack the origins of classroom practices," followed by a series of high-focused interviews. The purpose was to find out if the high-stakes assessment significantly affected practices at school even though the consequences were six years later, he said.

The researchers found that the 7th-grade forms absolutely replicated the 12th-grade assessment at the level of instruction and task. The conclusion, Clarke said, is that assessment is not neutral, "but a powerful component" of instruction and can be used as a "powerful and cost-efficient way to improve the system."

Pamela Aschbacher, formerly at CRESST and now at the California Institute of Technology, used portfolios for program evaluation to not only tell what students were doing but also to reveal what teachers were assigning from one school to another. The project was not looking at numbers as much as it was looking for teacher strategies that affected the cognitive knowledge of students. It is hoped that eventually the process will lead to self-evaluation and more effective professional development.

Working in eight schools in three districts and three classrooms in each school, the researchers asked the teachers for a collection of their assignments—such as typical homework, writing assignments, a challenging major project in language arts—and for each assignment, two high-level and two medium-level examples of student work. Two raters established rubrics for looking at the student work and the assignments, determining the cognitive demands of tasks and how well each task aligned with the teacher's goals for the students.

"The interrater reliabilities were not dynamite, but made a good start," Aschbacher said, noting that it was difficult to give a high rating on grading criteria when a teacher's goals were "incredibly vague." However, this kind of evaluation can lead to the development of a learning environment profile that describes the intellectual challenges of assignments and their alignment. One of 6 elementary teachers had both challenging assignments and good alignment, while only 1 of 10 middle school teachers did. Also, evaluation of the student work piece can "start conversations on what we expect students to learn and how the assignment relates to that."

In another study, Lindsay Clare of CRESST/UCLA observed classrooms and collected certain student assignments to determine teaching quality. The quality of

the instruction was determined by several factors: academic intensity of the curriculum; clear articulation of goals focused on learning; alignment of practice, assignments, and assessment; instructional feedback from teachers; and classroom management. The study used the NAEP 6-point scale for assessments as well as the criteria developed by Aschbacher.

There was good reliability in the ratings of teachers' assignments, Clare said, after a lot of training and a year's experience with the project. "There definitely is an overlap between what students learn in the classroom and teacher assignments, but it depends on what you want to measure," Clare said. The study eliminated homework assignments as a data source "because they were all over the map."

It is critical, Clare said, "for teachers to have a safe environment in which to share work, such as the Critical Friends strategy under the Annenberg Challenge." In this process, teachers can specify what they want others to look at. Having external facilitators helps because if the project is a function of the district, teachers always will think the evaluation is going directly to personnel.

Even as outsiders, "we were considered with suspicion, so you need to have the same people working with the teachers. Using student work is labor intensive, and part of the problem is that teachers don't have support such as secretaries, so we try to make it as easy as possible for them." Teachers are independent contractors in a sense, Clare said, and "once you are behind the classroom door, who wants someone breathing down your neck. The only way to be accepted is to make sure teachers know you are really there for positive purposes, that they see you as helping them facilitate learning."

Are We There? Issues in Testing and Validity

At the end of two days filled with descriptions of the aspects and pitfalls of complex assessment systems, Robert Mislevy of CRESST/Educational Testing Service began the closing session of the conference with a discussion of how to make sense of the data from such systems. The opportunity to do so is one of timing, he said, because educators and researchers can capitalize on cognitive educational psychology, or how people learn, and the benefit of new technologies.

Making sense of complex data, according to Mislevy, is really a task of reasoning in terms of stories, "weaving some sensible story around the specifics."

Science, he noted, is basically about checking out stories. He discussed several models for an assessment design, including

- the task model ("tasks are the most visible element in an educational assessment");
- the student model, or what complex of knowledge, skills or other attributes should be assessed presumably because they are tied to explicit or implicit objectives of instruction; and
- the evidence-centered model, which assesses the behaviors or performances that reveal student skills and knowledge, the observable variables.

Mislevy used a computer-based simulated performance assessment of problem solving developed for the Dental Interactive Simulation Corporation to illustrate how an assessment can tap knowledge and elicit behaviors that show this knowledge. Such assessments are called for in today's environment where there is a demand for more complex inferences about students. The implications for the student model from cognitive task analysis in simulations are that variables should be consistent with results of the task analysis and the purpose of the assessment. The implications for evidence models are that cognitive task analysis produced performance features that characterized recurring patterns of behavior. The implications for the task model are that developers can build cases around the features of problems that elicit evidence about the targeted knowledge and skills.

"The payoff from using a formal design framework," Mislevy said, "is that unique, complex cases can be created efficiently, along with an ability to score each one well."

Assessment, continued Robert Glaser of CRESST/University of Pittsburgh, "should reflect what modern knowledge of human cognition and mental processes tells us about human behavior and use of knowledge." The two fields of human cognition and psychometric aspects of assessment have come a long way in learning how to talk to each other, but Glaser would not say "we are there" yet.

The classical idea that education consists of study rather than just instruction became lost when 17th-century ideas of didactic teaching became dominant, "and the construct of student control over their learning became obscured," Glaser noted. The classical idea is coming back, he predicted, "as many investigators inform us that competence is fostered through teaching that engenders specific kinds of

cognitive activity." Structured knowledge "is not a consequence of the amount of knowledge received, but reflects exposure to an environment for learning where there are opportunities for problem solving, analyzing, making interpretations, and working in unfamiliar environments requiring transfer. These are significant targets for assessment."

Conditions of cognition, in which learning is a process of constructing new knowledge on the basis of current knowledge, are evident in three ways, according to Glaser. The first is representation and organization of knowledge. The way students represent information given depends on their organization of existing knowledge. Such structures enable individuals to build a mental model, thus helping them avoid trial and error, and to build structures for new learning and problem solving that go beyond surface features. As contrasted to new learners, advanced learners organize information around principles and practice. Knowledge must be acquired in such a way that it is highly articulated and integrated. The organization of knowledge, Glaser said, "provides a strong basis for thinking and for activity."

Another way that learning is evident is through self-regulation and self-agency. With increasing abilities, a learner can generate a representation of situations where information can be integrated with process. This enables advanced learners to leave their teachers behind, Glaser said. In this context, the skills of self-management become significant. The use of situations for learning varies with the friendliness of the environment. Assessment focuses on advanced learners' abilities.

Finally, social and situational accordances make learning evident. In a responsive social situation, learners can adopt what they see in others. They observe how others reason through group efforts. Students develop a facility for giving and accepting help and stimulation from others. Assessment requires performance in group activities.

Glaser called for better understanding of the kind of learning that fosters connected knowledge, noting that multiple-choice assessments poorly integrate knowledge. "The mode of research and development is changing," he said. "In the past, much of our attention was spent in moving from basic research to application, but in the future, we can best contribute to science and practice if we take on significant applications and ask how they correct and generate new frameworks for design situations that promote active knowledge."

Summing up the conference, Edmund Gordon of Yale University noted that the proceedings had "underscored the complexities of the problems we are dealing with, especially how interrelated and interdependent the issues are." Even more problematic, he said, was the question the conference was asked to address: "Are we there?"

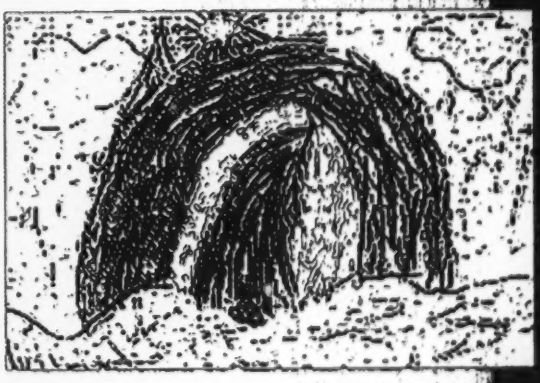
The extent of awareness about the problems is a product of experience with the phenomena one is trying to understand. Radical empiricism, he pointed out, contends that senses collect data, and if they are consistent with experience and hypothesized relationships, "we think we have good inferences. But how reliable are human experiences and instruments based on them?" One must ask, he said, "Do the data make sense in terms of what we already know about the real thing?" Problems arise when "we look for consistency in the evidence, but the evidence is the product of variance in human perceptions and conceptions of what is being observed." The more adjustments that are made to ensure reliability, "the further we may get from validity." Noting that Mislevy focused on the mechanism of assessment and Glaser on conceptual frameworks, Gordon said that "neither dealt with meanings of the behaviors. We must make judgments about what these things mean."

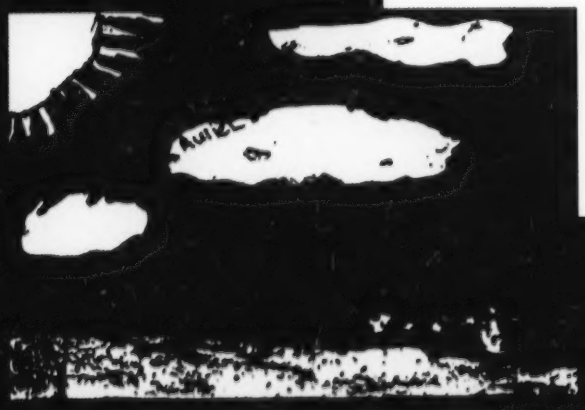
According to Gordon's calculations "we simply cannot get there. Issues of validity and testing challenge the very foundation of assessment. At best, we can only move toward valid measures. And finally, I have serious doubts about our trying to get there."

Reporting on small-group sessions that focused on future research on accountability, Eva Baker cited three major themes: equity and access; capacity building to interpret information, help children learn, and help communities understand what assessments are trying to accomplish; and support for multiple measures to broaden ways of understanding how and what children have learned. The groups, however, could cite few examples of where people were confident the accountability work was of high quality and could be pointed to as a model.

CRESST would continue to try to be better at communicating "what we think is important and what our constituencies think is important about student learning," Baker said. "The media often don't get it right, nor do we." The organization also will pursue system validity in all its complexities and difficulty—that is, systems of information, not a simple test. One thing that differentiates the work today from

where it started 20 years ago is that it is not laboratory work anymore, but is generated in schools with real time lines and people to be served. The focus for CRESST, she confirmed, would continue to be quality, equity, and usefulness of research in education.





UCLA Graduate School of Education & Information Studies